AD_____

DEVELOPMENT OF A TOXIN KNOWLEDGE SYSTEM

Annual Summary Report

Harold L. Trammel, Pharm.D.

May 15, 1988

for the Period April 6, 1987 through April 5, 1988

Supported by

U. S. ARMY MEDICAL RESEARCH AND DEVELOPMENT COMMAND
Fort Detrick, Frederick, Maryland 21701-5012

Contract No. DAMD17-87-C-7114

Department of Veterinary Biosciences
College of Veterinary Medicine
University of Illinois
Urbana, Illinois 61801

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: University of Illinois College of Veterinary Medicine; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code): Department of Veterinary Biosciences Urbara, Illinois 61801	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Medical Research & Development Command; 8b. OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: Contract No. DAMD17-87-C-7114
8c. ADDRESS (City, State, and ZIP Code): Fort Detrick Frederick, Maryland 21701-5012	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 61102A; PROJECT NO. 3M1-61102BS12; TASK NO. AD; WORK UNIT ACCESSION NO. 096

11 TITLE (Include Security Classification): Development of a Toxin Knowledge System

12. PERSONAL AUTHOR(S): Harold L. Trammel

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Annual Report	FROM 4/6/87 TO 4/5/88	1988 May 15	51

16. SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Artificial Intelligence; BW; Database; Assessment; Toxins Monograph
06	11		
06	15		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

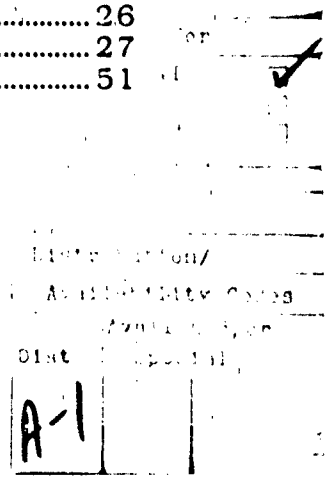
The Toxin Knowledge System (TKS) is being developed to provide rapidly accessible, up-to-date knowledge about low molecular weight toxins. This system will integrate facts from published literature into monographs on individual toxins. Development utilizes a relational database management system and associated fourth-generation programming language on minicomputer. It uses a standard knowledge structure, structured abstracting processes, standard nomenclature systems, and computer-generated structured monographs. This system exploits the structured style of scientific writing to collect information on low molecular weight toxins and store this information in structured form. The application guides the abstractor in this collection process, facilitating the extraction of similar information from different papers. A sophisticated user interface allows the user to readily add, find, delete, and update data in the system. Currently TKS manages citation data for both journals and books, has keyword access to entered citations, and can collect information on a paper including the study designs, the subjects and exposure regimens used, and the links to connect this data to the clinical findings reported. Controlled vocabularies are used for journal titles and abbreviations, book titles, and keywords. Future efforts will stress incorporation of clinical finding data, analytical methods and results, mechanisms of action, and pharmacokinetics. Monograph generation programs will be written to extract and compile clinical and pathological findings reported with a subject group exposed to a particular toxin.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Mrs. Virginia M. Miller	301-663-7325	SGRD-RMI-S

DD Form 1473, JUN 86 — Previous editions are obsolete — SECURITY CLASSIFICATION OF THIS PAGE

Table of Contents



List of Figures

List of Appendices

SUMMARY

In order to meet the need for rapidly accessible, up-to-date knowledge about low molecular weight toxins, a Toxin Knowledge System was initiated. The current information tools such as citation indexes and abstracting systems do not integrate new facts into an existing knowledge base. These tools provide only citations or narrative abstracts to published papers. Development has begun on a Toxin Knowledge System which, when completed, will integrate facts from published literature into a readily useable monograph on individual toxins.

The Toxin Knowledge System is being developed on a minicomputer using a relational database management system and associated fourth-generation programming language. It uses a standard knowledge structure, structured abstracting processes, standard nomenclature systems, and computer-generated structured monographs. This system exploits the structured style of scientific writing to collect information on low molecular weight toxins and store the collected information in structured form. A structured abstracting technique is used to guide the abstractor in this collection process. Structured abstracting requires the answering of a standard set of questions about the content of the paper, thereby facilitating the extraction of similar information from different papers. The goal of this system is to prepare continuously updated monographs on toxins as new papers are processed.

The use of a fourth-generation computer language has permitted the creation of a sophisticated user interface for data manipulation. This interface uses windowing, menus, dialog boxes, scrolling arrays and dynamic on-screen displays of possible options. The user can readily add, find, delete, and update data in the system. The current version of the Toxin Knowledge System can manage the citation data for both journals and books in an similar manner, has keyword access to entered citations, and can collect information on a paper. This paper information includes the study designs used in the paper, the subjects and exposure regimens used in the designs, and generate the links needed to connect this data to the clinical findings reported in the paper. Controlled vocabularies have been created for journal titles and abbreviations, book titles, and keywords. Additional controlled vocabularies are being developed for clinical findings (based on *SNOMED/SNOVET*) and generic agents (using *RTECS* and *USAN*).

When completed the system will be able to extract a detailed set of clinical and pathological findings reported with a subject group exposed to a particular toxin. These clinical findings will be sorted by body system, organ, and finding. Findings from one paper will be presented in conjunction with similar findings from other papers. Treatments reported in published papers will also be compiled and should give insight into which treatments are most effective.

The system development has progressed significantly but is incomplete. Additional database tables are needed to collect data on analytical methods, analytical results, mechanisms of action, and pharmacokinetics. The structured monograph generation needs to be more fully developed. When the necessary tables are in place, the abstracting process will be stressed and monograph generation will be further reviewed for correctness and clarity.

FOREWORD

Citations of commercial organizations or trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

I. STATEMENT OF PROBLEM

I.A. Need for Knowledge About Low Molecular Weight Toxins

Military and civil defense authorities are concerned about the production and use of toxins against military and civilian populations. If an enemy were to use a toxin in an attack on military personnel, it would be imperative that the toxin be detected, a diagnosis made, and appropriate treatment implemented rapidly to decrease the adverse effects of the attack.

While there is a growing body of information about toxins, there has been no system to collect and compile this information into a readily usable knowledge system. The specific needs for knowledge about toxins varies with the user. Researchers studying toxins need detailed, current reference information compiled from both the literature and other research groups. Military and civil defense health professionals need ready access to extensive information on the detection, diagnosis, and treatment of medical problems associated with toxins. Military personnel in areas where exposure to toxins is possible need immediately available references that are current and appropriate for the individual's training and situation.

While the knowledge each group needs is varied, the factual basis for this knowledge is derived from the same literature sources. All groups would benefit if there were an efficient means to collect toxin information in such a manner that the knowledge needs of each group can be met from a single source. This source of toxin information or knowledge should consist of detailed, comprehensive information, but be able to provide each group with the specific facts and details appropriate for the needs of the group.

I.B. Problem of Maintaining Knowledge

Scientific knowledge can be defined as the sum total of what is known about a topic or as a body of systemized facts, information, principles, and experiences relating to a singular topic. Gaining this knowledge requires collecting information about the topic and compiling that information into a readily usable form. This is a difficult and time consuming process. Keeping knowledge current is even more difficult. As research uncovers new facts about the topic, they must be incorporated into what is already known.

To date, efforts to meet the need for current information have generally failed to incorporate new facts into a usable form. The two most common means of providing access to current literature are citation indexes and abstracting systems.

Citation indexes provide only citation information for pertinent literature sources. The purpose of citation indexes is to provide users with journal article citations from which the original article can be obtained. A user wanting to gain information from a citation index would use some form of keyword-based search strategy to find the desired literature citations. The user would then have to find the actual article in order to obtain the facts necessary to add to his/her knowledge. Citation indexes continue to be important ways to access the published literat re. This form of system is the foundation of most other information systems. A major problem with using only citation indexes for gaining knowledge is that the information provided is simply a pointer to the facts and not the facts themselves.

Abstracting systems start with the citation index foundation and add narrative abstracts of the paper. The abstracts are used to improve the efficiency of selecting journal articles for detailed review. The abstracts in these systems can provide facts which increase knowledge on the topic. The amount of scientific information contained in these abstracts is limited in part by the narrative format of the abstract which necessitates the facts being contained in a sentence format. Usually the user will need to obtain and review the original paper to gain the knowledge s/he needs.

There are two major difficulties with using either citation indexes or abstracting services as a source of knowledge. The first and most important is the time needed to gain the needed information. The system must be searched and appropriate titles identified. Both methods require obtaining the original article to find desired facts. This entails finding the article and either reading it in the library setting or copying it for later reading. The user must read the paper, take notes, and attempt to synthesize an overview of all the articles and their content. This synthesized understanding of the literature is knowledge.

The second major difficulty is the need to keep this knowledge up to date. As more scientific information is published on a topic, it needs to be incorporated into the previously synthesized understanding. The process is compounded by the usual need to review the previously obtained papers while reading the new papers in order to see how the new data fits together with the old. From this, a new understanding is reached and knowledge is now updated. This is a time consuming process.

II. APPROACH TO PROBLEM

Our group has extensive and varied experience in the preparation and delivery of biomedical information. For several years, we have routinely provided answers to specific toxicologic and drug-oriented questions received from a wide range of individuals. We also prepare monographs and review papers about various toxicological or pharmaceutical agents for both internal use and for publication. Our toxicology research programs require access, utilization, and summarization of detailed information.

Using this background, we compared the knowledge acquisition techniques used by different individuals. This analysis revealed common methods and procedures as well as commonly accepted needs for how the knowledge should be made available.

II.A. The Knowledge Acquisition Process

Almost unconsciously, a scientist acquiring knowledge from the literature takes advantage of certain standard structures and terminologies. In using citation indexes and/or abstracting systems, s/he selects papers based on a standard keyword vocabulary. S/he uses the standard citation structure to identify the papers to be reviewed. This structure includes both the format of the citation and the standard abbreviations used.

After obtaining the desired papers, the scientist begins to read the papers and thereby uses the format used to write scientific papers. Each discipline has its own particular format and most papers from a given discipline are prepared according to that format. The standard format facilitates the scientist's

identification of the critical components of the study design and the associated results and conclusions.

Frequently the scientist will sort the papers by the study design used. S/he may group the papers by case reports and animal studies. From this sorting, the scientist may further group the papers by the materials and methods. For example, s/he might group papers by dosage regimens to consider them from a dose-response perspective. The reader may have to sort the papers several times in various ways in order to obtain an understanding of the study and its results.

When the authors of a scientific paper wrote the paper, their goal was to communicate how their work was performed and what their results were. They used "standard" terms in order to assure that the reader would understand what they did and saw. This is especially true with clinical findings seen as result of the study. If the scientist reading the paper is unfamiliar with a particular term, s/he must either "translate" it into a term s/he already knows or add this term to his/her vocabulary. Subsequently the reviewer will consider the author's discussion of results. Frequently the discussion in current papers will provide both a reference to and an evaluation of older papers.

The data from the individual papers must be integrated into a cohesive form by the scientist. The result data from the various papers are considered by the reviewer as groups of results, along with the study design, materials and methods used, and conclusions drawn from the results. The form that the scientist's summary may take is quite varied. The end result can be a printed monograph on the topic, or may be kept only in the mind of the scientist.

Unfortunately, textual materials, such as reference books, monographs, and text books, are frequently neglected in this process. Too often, the scientist seeks his/her answers only in current literature with limited success, and yet part or all of the answers may have been published several years earlier and summarized in textual materials. Many times these important sources of information yield a deeper understanding, especially with regard to the historical development of an idea or procedure. This information should be able to be included with current journal articles to provide a more comprehensive understanding.

II.B. Automation of the Process

We believed this process could be automated to a significant degree. While we would not expect an automated system to be able to write a paper for publication, we believed that by mimicking the knowledge acquisition process and by utilizing the inherent structure of the literature we could develop a systemized method to extract needed data about toxins and compile that data into usable, continuously updated knowledge.

This method would be based on four elements:

1) a standard knowledge structure
2) a structured abstracting process
3) a standard nomenclature system
4) a structured monograph design.

By predefining the structure and terminology, the individual pieces of data from scientific papers could be collected into a composite knowledge source which can be readily accessed for answers to questions.

II.C. A Standard Knowledge Structure for Toxin Information

The **standard knowledge structure** we envisioned would model biomedical literature. We initially conceived the structure to consist of the following:

Initial Toxin Knowledge System Structure
Citation data
Author data
Article Type data
Study Design data
Subject data
Exposure data
Pathophysiology data (System, Organ, Finding)
Chemical data
Results data
Management data
Critique data

Having a standard structure for the data will of necessity lead to ordered collections of facts. The benefits of a standardized structure include consistency throughout the system and facilitating the identification of missing data which may mean research needs to be performed. The major problem with standardized structures is the exceptional paper that will not easily fit into the structure. We believe that the benefits outweigh the problems and that with work, the structure can include more of the exceptions.

II.D. Collect Data Using a Structured Abstracting Approach

To obtain the data from the published source, we proposed the use of a **structured abstracting** approach. Structured abstracts differ from the traditional narrative abstracts in that a predefined structure is used to present information from a report and unnecessary prose is avoided. Similarly structured abstracting uses a predefined structure to obtain the information published in the report. By having a standard mechanism to extract information from the papers, more comprehensive data collection is likely.

II.E. Standardized Nomenclature

The use of controlled vocabularies is incumbent in order to provide consistency in the terminology used. This is especially true if the data from many different papers are to be compiled into one knowledge set. We initially identified two areas where a controlled vocabulary would be essential. These were generic agent names and clinical finding terms. The generic agent names would include the preferred names of toxins.

II.F. Compile collected data into Structured Monograph

The eventual output of the proposed knowledge system was **structured monographs** on diagnosis and treatment. These two monographs would use a structured monograph technique to present the information collected in the system. The structured nature of the abstracting process and storage in the database would be exploited to produce a monograph with the data collected into a

standard structure. The use of structured abstracts to represent knowledge is consistent with the definition of knowledge as ordered sets of facts. The fixed structure provides an ordered means for the facts collected into the system to be presented in such a manner that the information can be quickly found. These facts would of necessity have an indication of the factors influencing them. For example, the dose of a toxin required to produce a given clinical effect must be presented with the clinical effect to give a true representation of the facts.

II.G. Computerized Methodology

To make the standard toxin knowledge structure workable, we proposed using the Informix-SQL™[1] relational database management system on a minicomputer. As an abstractor read a journal or textual information source, s/he would interact with the database program via a computer terminal. The program would present questions and prompts to be completed by the abstractor using data from the papers. The data would be stored in various database tables and would be linked via a unique citation number; thus, all entries for a given paper or book would be extractable as a unit of information.

Our efforts in designing a comprehensive veterinary toxicology case record database indicated that in order to get the level of detail and accuracy needed to fully describe an article, the abstracting process would need a well-conceived user interface with on-line checks for data consistency. The user should be able to flow through the abstracting process smoothly. S/he should generally be able to read a given paper and easily enter the data from it. Critical key-field data should be generated automatically if possible. The user should be able to see what options exist at any point in the process and should be able to look-up possible entries with limited effort. Varying degrees of user experience would have to be considered when designing the interface.

Data entry systems which require the user to enter the links between the various interactive database tables are prone to mistakes. The underlying processes to maintain the database, such as links between tables, should be somewhat hidden from the user. Data manipulation should be done via the interface instead of direct user interaction with the data in the tables.

We believe that both journal and book data should be included in the system and be managed in a similar manner. Our group's earlier experience in developing a small bibliographic system suggested that the apparent differences in citation styles would require a separate citation entry process for each. To have separate methods to handle book and journal data would be opposed to the basic design of our proposed system; thus, some means would have to be developed to manage the apparent differences.

III. RESULTS

III.A. Database Software

We began developing the Toxin Knowledge System with Informix-SQL™ relational database software on a Sequent™[2] minicomputer. We had had extensive experience with this software on a small multi-user computer and

---

[1]Informix Software, Inc , 4100 Bohannon Drive, Menlo Park, California 94025
[2]Sequent Computer Systems, Inc., 15450 S.W Koll Parkway, Beaverton, Oregon 97006-6063

found it to be an excellent relational database package. We were able to begin creating the various database tables with rather complex interactions in a short period of time. As we began to test the initial table design by entering data, we found limitations with the data entry process. Informix-SQL™ has a good screen entry program, but we found that this program would not let us create the user interface we had in mind. This program could not hide many of the complex interactions between the various data tables, and would not permit implementation of the user interface we had begun to realize we needed.

Some of these problems had been anticipated and we had originally proposed to use Informix-ESQL/C™ to provide additional needed features we believed Informix-SQL™ needed. Discussions with Informix technical representatives led us to conclude that we could more quickly create the interface with Informix-4GL™, a fourth-generation computer language for Informix-based databases. This would permit us to utilize the strengths of Informix-SQL™ for the general database management process and have essentially full control over the user interface design. We elected to take this course of action even though it would require our learning a new programming language. We have not regretted this decision, because Informix-4GL™ is a powerful language with a wide range of features, and is not limited to use with a database.

Until the Informix-4GL™ based program was developed, we continued to use the Informix-SQL™ based entry methods to enter citation data for toxin-related journal articles previously collected by members of our group. The use of these methods revealed the areas of interaction that needed to be managed by the Informix-4GL™ program, rather than relying on the user. It also identified the need for a controlled vocabulary for journal abbreviations and for providing better access to the collected data.

III.B. Citation Data Processing

The foundation of any knowledge system using the published literature is the citation. If we were to effectively extract data and subsequently construct a monograph, we needed to insure that the citation data for a given paper was collected, stored, and made accessible in an optimal fashion. This component of the Toxin Knowledge System became a keystone in the development process for four reasons:

1. The citation data itself was important in the overall design.
2. The Toxin Knowledge System should handle both journal and book information equally well. The disparate style of citations for books and journals had to be overcome.
3. We needed to learn Informix-4GL™ programming techniques and this provided a reasonably well-defined section for use in developing the initial user interface program. We had experience in using Informix-SQL™ to enter this data and thus had a clear idea of what the finished module should do.
4. Controlled vocabularies were needed for both journal and book titles. These vocabularies needed to be available on-line to the user. The programming techniques needed to provide this would be used extensively in other sections.

The citation processing module was successfully developed in accordance with our underlying design for both the database tables and the user interface. The primary design problem for this module was the above-mentioned style

differences for book and journal citations. To resolve this, we compared the elements of both citation styles and identified the elements that were common. Journal articles and book chapters contain many of the same elements; however, the citation for the book containing the chapter has many unique elements. We had determined that the full journal title would not be used in this module and, at most, journal abbreviations would be used. To reduce the amount of typing needed, we thought that a code to the journal would be better. Our analysis resulted in the following:

Journal Citations	Book Citations
Authors (many)	Chapter Authors (many)
Article Title	Chapter Title
*Journal Reference**	*Book Reference†*
Journal Volume	Book Chapter Number
Journal Pages	Chapter Pages
Year	Year

* Journal Reference	† Book Reference
	Editors
Journal Title	Book Title
(Journal Abbreviation)	Edition Number
	Volume Number
	Edition Date
	Publisher
	Place of Publication

Four database tables were created to hold the various elements of the citation data. The contents of these tables are presented in Appendix A, and their interactions are depicted in Appendix B.

III.B.1 Journal Vocabulary Table

A table to hold the journal reference data was created. This table, **journlst**, serves as a journal name controlled vocabulary. Each journal title was assigned a code number consisting of the letter **J** followed by a sequentially assigned accession number. This *jcode* is used as a link to the **citation** table. The journal title and abbreviation used was usually consistent with the National Library of Medicine (NLM) *List of Journals Indexed*. Many journals require that authors use the NLM abbreviations, and we decided to adhere to this *ad hoc* standard. Abbreviations and titles for journals not found in this list were taken from the journals themselves. Journal names and abbreviations can be added to the vocabulary as needed, even while the user is putting journal citation data into the computer.

Our initial efforts in Informix-4GL™ programming were aimed at developing a program module to manage the journal vocabulary data. This program automatically assigns the sequential accession number and generates the *jcode* value for any new journal added to the vocabulary. The user can search for any item in the journal vocabulary and update or delete it as is needed.

III.B.2 Book Vocabulary Table

Similarly, a table to hold book reference data was created and an Informix-4GL™ program module prepared to manage this table. **Booklst** contains all the elements necessary to identify the specific book. Each book is assigned a *bcode* number consisting of the letter **B** and a sequentially assigned accession number. Like the *jcode* in the journal vocabulary table, this value is used to link the book data to the **citation** table. Book data is entered as needed and can be added while the user is entering book citation data.

III.B.3 Citation Table

We decided that the similar elements of the journal and book citation could become the identifying data for a given citation. We put these elements into the **citation** table. The **citation** table would serve as the master table for all subsequent data tables. The citation would link to the journal or book reference via the *citsource* column. Other data tables would link to the **citation** table using a citation code number created when the citation is first entered into the Toxin Knowledge System.

The user interface for this module uses a mixture of menus, screens, prompts, and dialog boxes. The data entry screen for this table is shown below in Figure 1.

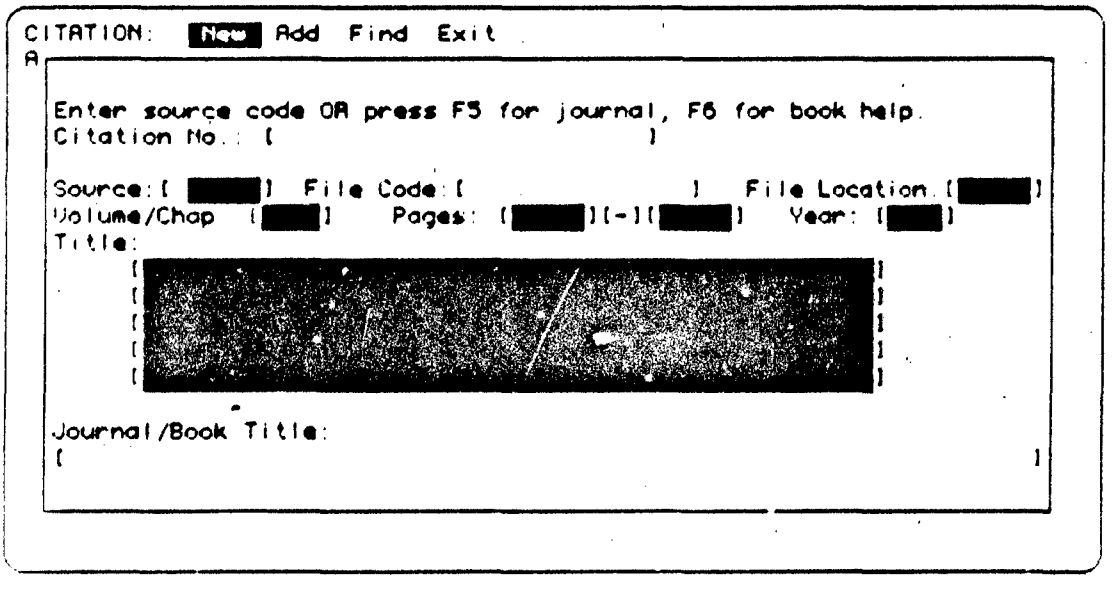


Figure 1. Citation Entry Screen

When the user first accesses this screen, the cursor is in the Citation Source field and a message indicates that journal sources and book sources are available for look-up at the press of a function key. Figure 2 shows an example of the journal look-up screen. Depending on the function key selected, the user can query for a journal abbreviation or book title using wildcard searching. Up to thirty entries meeting the search criteria are displayed in the window. The user can scroll through these entries and select the desired journal or book by pressing the Escape key. The look-up window disappears, the selected journal or book code

is automatically inserted into the Citation Source field, and the journal abbreviation or book title is displayed for verification. The user can elect to change this entry by entering a different number or pressing the look-up function key again.

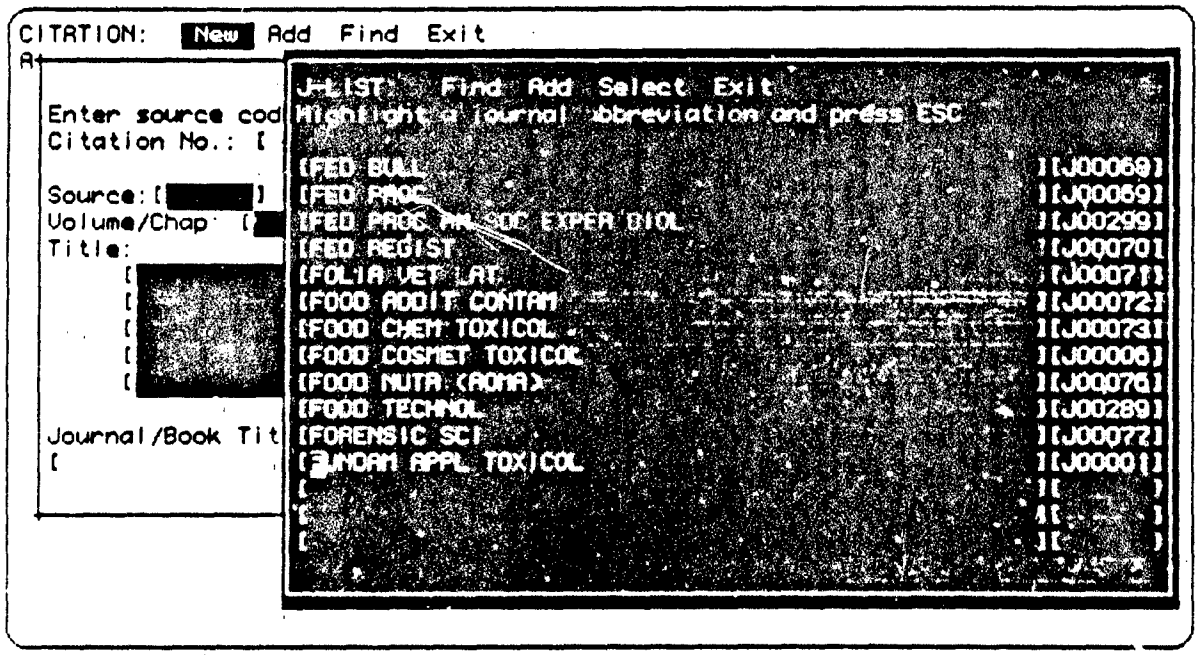


Figure 2. Citation Screen with Journal Look-up Window

After the Citation Source entry, the user continues to enter appropriate data into the screen entries. After year value is entered, the program automatically generates the citation code number and puts this in the corresponding field. This code number is composed from the Citation Source value, the volume/chapter number, the first page number, and the year. For example, a citation from *Furdamental and Applied Toxicology*, volume 9, pages 1588 to 594, published in 1987 would have the following citation code number: **J00001-0009-00588-1987**. When all of the appropriate data is entered, s/he pushes the Escape key, the data is inserted into the **citation** table in the database, and the author entry portion is called.

III.B.4 Author Table

Because the number of authors varies, we used a separate table to hold the author names and their order of authorship. Each entry was joined to the other tables via the citation code number. The entry screen for this table is shown in Figure 3 below. The citation code number is automatically displayed to assure correct links to the **citation** table. The user enters the authors' names into a scrolling entry array. Assuming the names are put into the system in order, the program will automatically generate the publication order number as the user puts additional names into the array. The current system allows up to 20 authors' names to be entered.

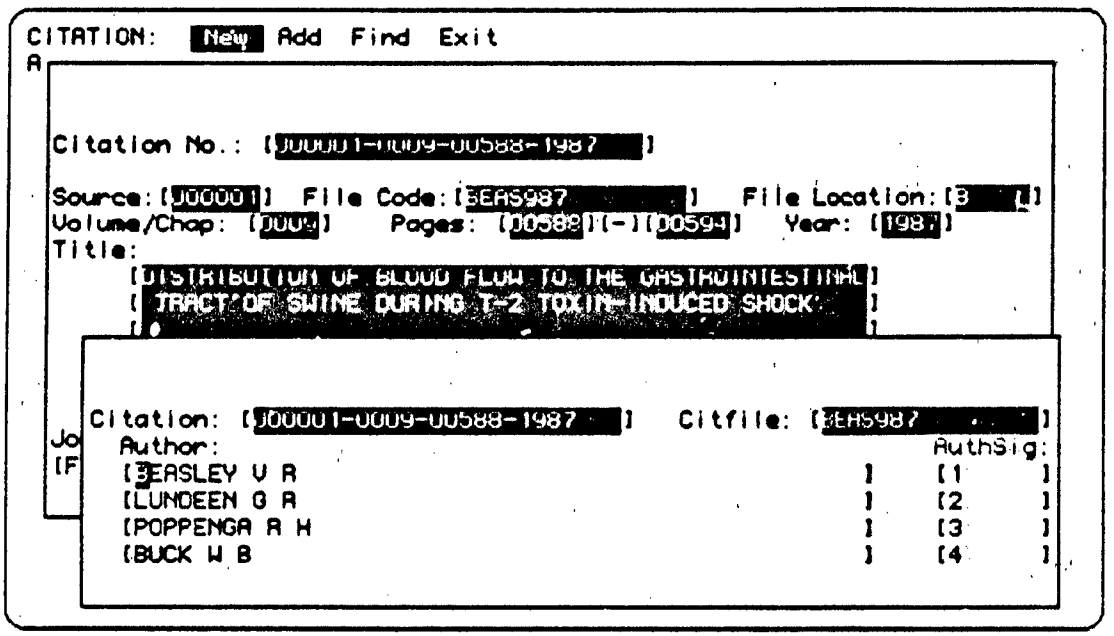


Figure 3. Author Entry Screen

III.C. Addition of Keywords to Toxin Knowledge System

Because of the entry of toxin-related articles prior to the completion of the Toxin Knowledge System, we decided to add a **keywords** table and associated **keylist** table to the system. Appendix C contains a description of the contents of the keyword-related database tables. If the Toxin Knowledge System were complete, these tables would not be necessary; however, in order to be able to access and select papers for full abstracting when the system is complete, we believe this addition is necessary at this stage. This also makes the system useful prior to completion.

As part of the Informix-4GL™ program to collect keyword data, we worked out the techniques necessary to have on-line checks for data correctness. The Informix-SQL™ entry method did not have this feature, and each user had the option to modify the keywords that were being used. The Toxin Knowledge System now has a list of accepted keywords in the **keylist** table which is used for verification and on-line look-up. The interactions between the **keywords** table and the **keylist** table are depicted in Appendix D.

When a new citation is entered into the system, the user will enter the citation and author data as described above. The keyword module is then activated to permit entry of up to 20 keywords. The screen used to enter this data is shown in Figure 4.

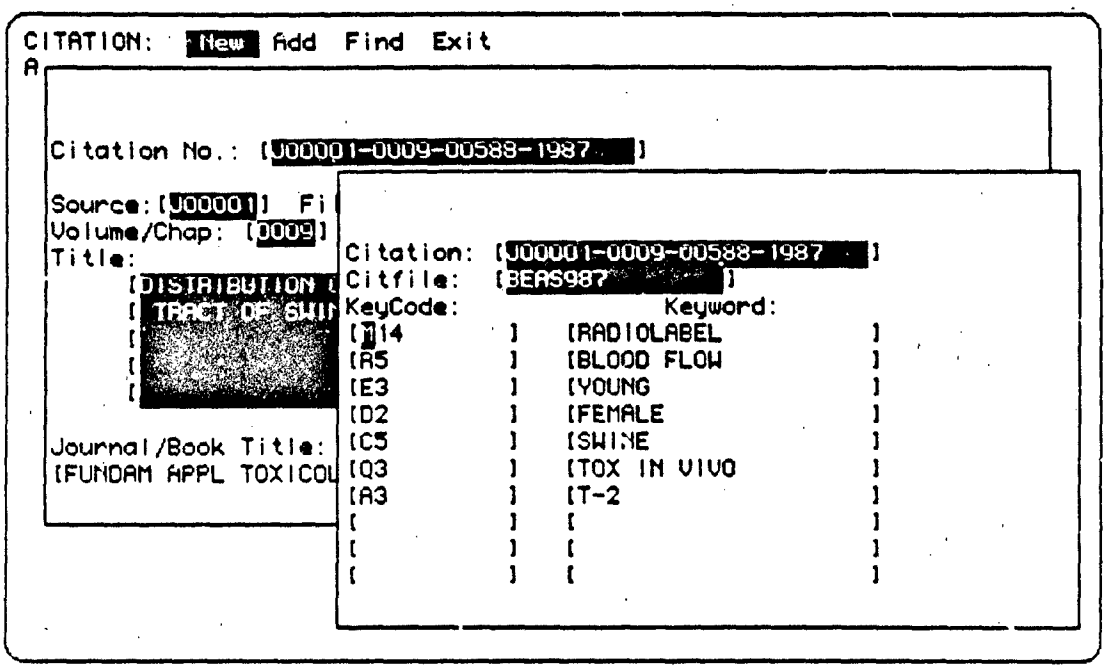


Figure 4. Keyword Entry Screen

Users can either input a code and the computer program will look up and insert the corresponding keyword or they can input a keyword and the corresponding code will be determined and inserted. This dual mechanism was found to be more effective than having only one mechanism. Users find that there are certain keywords that are frequently used. If they learn the code for these words, three keystrokes produce a keyword that would require up to 20 keystrokes. Infrequently used terms might be remembered as words but not as the associated codes. The current system addresses both situations.

III.D. Cross Table Query Process for Keywords and Citations

After the Informix-4GL™ program was developed for entry and retrieval of citations, authors, and keywords, we considered it essential that a mechanism be prepared to permit queries across all three tables simultaneously. We developed a query-by-example screen that would permit a user to enter search terms for any item in the citation table, up to three author names, three keywords, and four keycodes. Wildcard searches are supported in any field. This screen is shown in Figure 5 below.

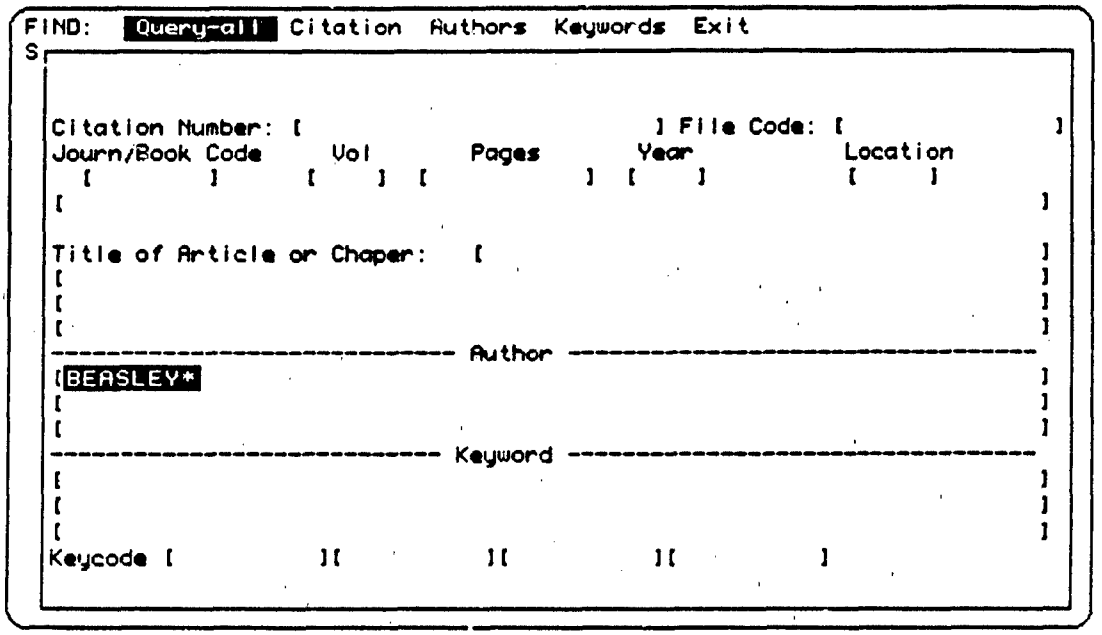


Figure 5. Query-by-Example Screen

The program queries for entries in all three tables which meet the appropriate search criteria. The program concatenates the author and keyword entries into character strings and displays them in the appropriate fields on screen. Figure 6 presents what the user might see on screen.

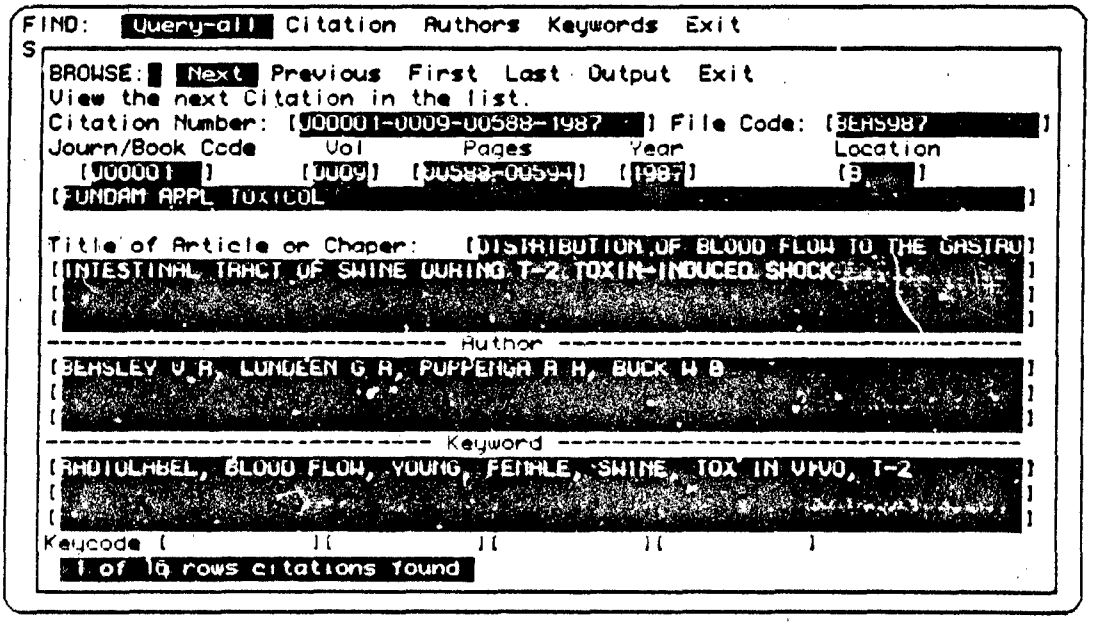


Figure 6. Query-by-Example Screen with Retrieved Data

Using the menus, the user can browse through the citations and elect to output all or selected citations to either a file or to a printer. The output strongly resembles a list of bibliographic citations sorted by the first author's last name. An example of such a citation output is:

```
TKS code: J00C01-0009-00588-1987
File code: BEA3987 in B files
BEASLEY V R, LUNDEEN G R, POPPENGA R H, BUCK W B: DISTRIBUTION OF
BLOOD FLOW TO THE GASTROINTESTINAL TRACT OF SWINE DURING T-2
TOXIN-INDUCED SHOCK, FUNDAM APPL TOXICOL 0009:00588-00594, 1987
Keywords: RADIOLABEL, BLOOD FLOW, YOUNG, FEMALE, SWINE, TOX IN
VIVO, T-2
```

This was our first effort using Informix-4GL™ to construct usable output from the individual facts stored in the system and we were pleased with how well it worked. We will build on these techniques extensively as we extend the Toxin Knowledge System.

III.E. Paper Data Processing

With the citation and keyword modules essentially complete, we turned our attention to processing the content of the papers. After a detailed analysis of representative papers being entered into the system, we altered the working components of the initial Toxin Knowledge System structure. The current working structure is presented below.

Revised Toxin Knowledge System Structure

- Citation data
  - Author data
- Keyword data
- Paper Overview Section
  - Article Type data
- Methods and Materials
  - Methods
    - Study Design data
    - Analytical Methods data
  - Materials
    - Subject data
    - Exposure data
- Results Section
  - Clinical Findings (Pathophysiology) data
  - Pharmacokinetics data
  - Chemical data
- Discussion and Comments Section
  - Critique data

III.E.1. Paper Overview Section

The paper overview section is the master section for all content sections in the Toxin Knowledge System. This section is made up of a single database table, **paperover**. The contents of this table are presented in Appendix E. Appendix F shows the interactions this table has with the tables in the Methods and Materials Section.

For each paper there is only one entry in the **paperover** table. It serves as a foundation for the multiple entities in the other content tables. In addition to the table-to-table linking information, this table contains certain basic information about the paper. Both the stated purpose of the paper and the abstractor's impression of an implied purpose are collected and stored here. An implied purpose can frequently give insight into the authors' biases that might be at work. This table also contains the aim of the paper. We have begun to establish a standardized list of acceptable terms for this item. We plan to eventually use this term as a controlling flag for the flow of the structured abstracting process. One such flag is the column for the number of study designs present in the paper. The abstractor indicates the number of designs at this point and controls how many study designs can be entered in the Materials and Methods Section. The data entry screen for this table is shown in Figure 7.

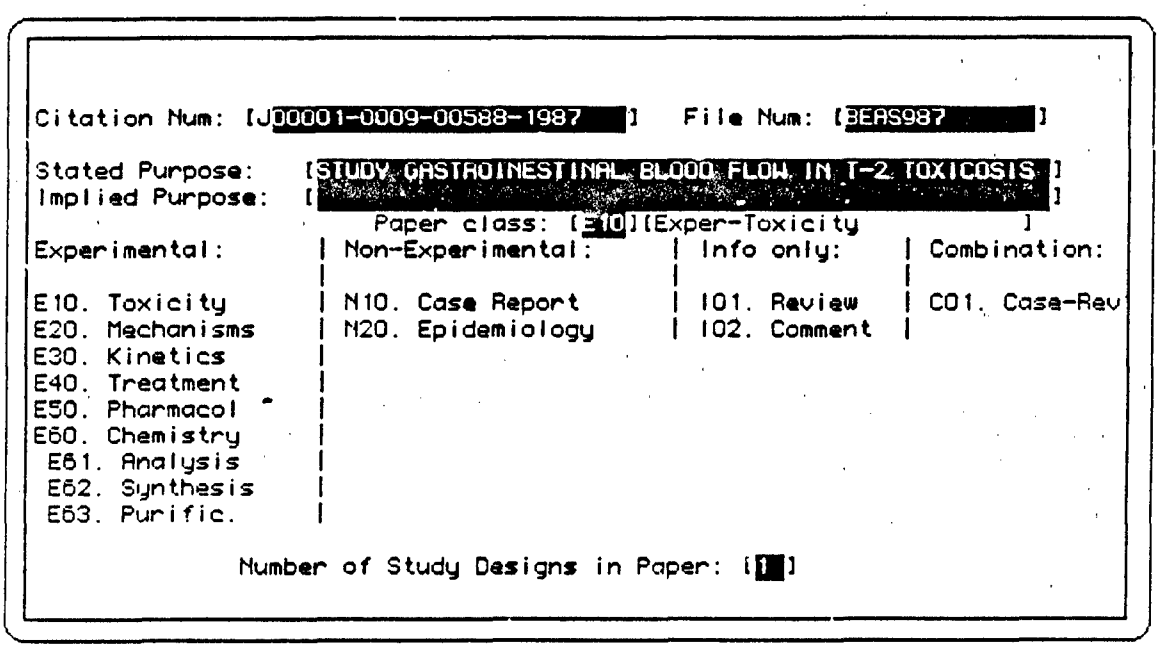


Figure 7. Paper Overview Data Entry Screen

The user's interaction with the above screen is generally straightforward. The citation number and file number are carried over from the citation entry process after a new citation is entered into the system. If the user intends to add content data for a citation already in the system, s/he will be prompted for the citation number. After the user indicates the citation number, this number and

corresponding file number will be put into the corresponding fields on screen. The user enters the purpose data and selects the desired aim or paper class from the choices available. After entry of the code number, the associated translation appears next to it. The user then enters the number of study designs in the paper. For example, if the paper consists of a case report of a human exposed to a toxin and an animal study to replicate the effects seen in the human, there would be two study designs and a 2 would be entered in the screen field. After all the data is entered, the user pushes the Escape key and the data is inserted into the database. If this is the entry of a new citation, the user will automatically go to the study design screen, or else the user is presented with the "Study-Methods — Materials — Results" menu.

III.E.2. Methods and Materials Section

In keeping with the standard style for writing scientific papers, the Methods and Materials section contains the tables necessary to hold data about the various methodologies and materials used in the study.

III.E.2.a. Methods

Currently only one methods table is defined, that being **stdydsgn**, the study design data. In general, this table contains the general design information, the controlling technique data, the number of subject groups involved in this design, and the number of exposure regimens in this design. This table is described in detail in Appendix E. This table is linked to the materials tables by means of the citation and design numbers. Each design in a paper is assigned a number as it is entered into the screen shown below.

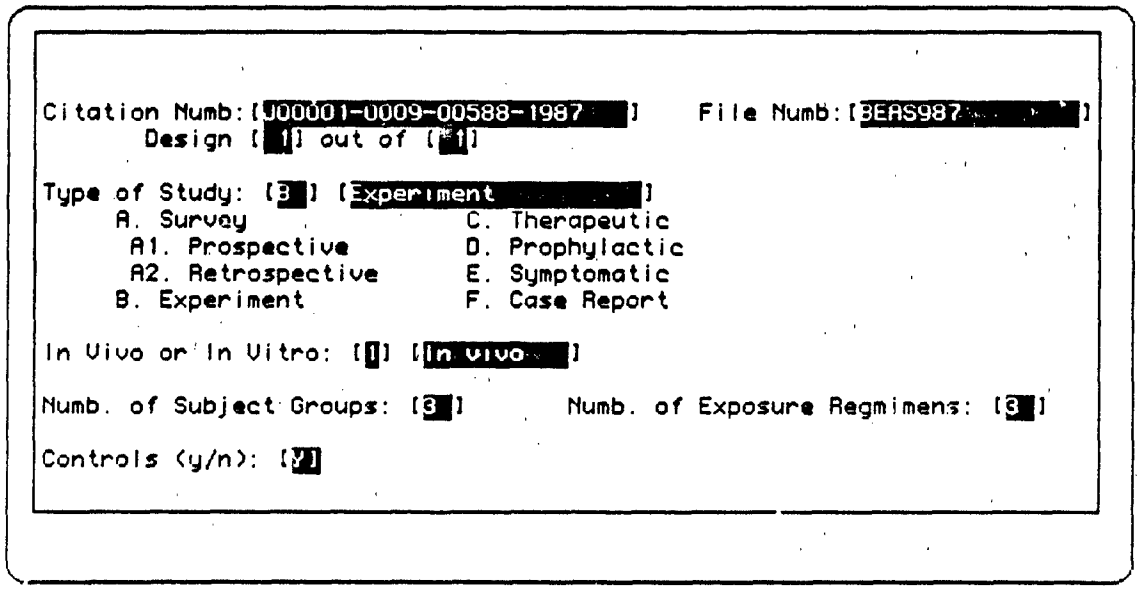


Figure 8. Study Design Entry Screen, part 1

Figure 8 shows the first of two data entry screens for study design data. Like the paper overview data, the citation number and file number data are

automatically inserted when the screen opens. The program also automatically maintains the current study design number. If a paper had two study designs, the program would show 1 out of 2 designs when the user was entering data for the first study design. The user would first indicate the type of study by selecting from the on-screen choices. The user would then enter a code number and the associated translation would appear next to it. Similarly, the user would indicate whether the study was an *in vivo* or *in vitro* study. The next item is the number of subject groups involved in this particular study design. This does not necessarily indicate the total number of subject groups involved in the entire paper. In like manner, the next field is the number of exposure regimens in this particular design. If the user enters a **Y** in the Controls field, the screen in Figure 9 will be displayed to input Control Technique data.

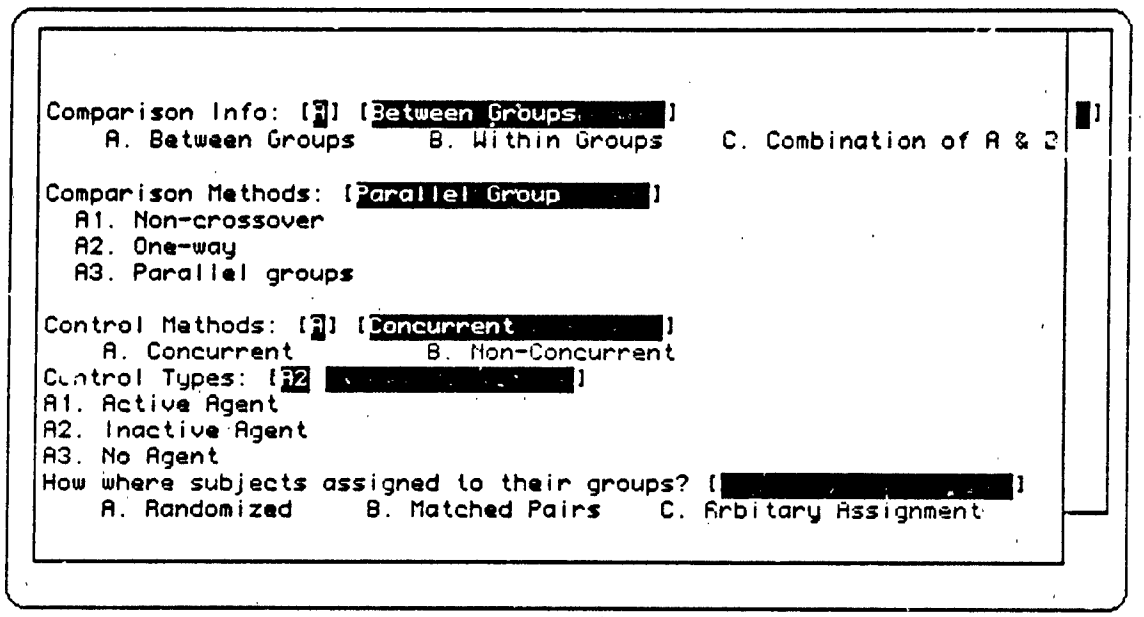


Figure 9. Study Design Data Entry Screen, part 2

This screen incorporates several user-oriented enhancements to direct the structured abstracting process. The user selects and enters the appropriate Comparison Information option and the corresponding translation will appear. In addition, the acceptable options for Comparison Methods will be displayed under the Comparison Methods field. In Figure 9, the user entered an **A** (Between Groups) in the Comparison Information field and the three choices **A1**, **A2**, and **A3** appear. If a **B** had been entered, different options would have appeared. The user had entered **A3** in the field and **Parallel Group** appeared in the data entry field by itself. This type of methodology is used for the other fields in this screen.

If the user pushes the Escape key and the correct number of designs have not been entered, the user will receive an error message and will be prompted to enter the remaining design data. After successful design data entry, the user will be returned to the "Study-Methods — Materials — Results" menu.

In the future, we plan to include an Analytical Methods table for data on the various methodologies used to detect toxins. This will be used both as a source of information on detection and quantification methods, and as a reference table for the results and clinical findings tables. When a paper reports the use of a particular methodology, that paper will be internally linked to the corresponding entry in this table.

III.E.2.b. Materials

When the user selects the Materials option on the menu, s/he sees a "Subject - Regimen - Links" menu. There are three database tables with materials data defined at this time. These are subject group data (**subjgrp**), exposure regimen data (**exporegm**), and a data table that holds the links between the different subject groups and the exposure regimens (**expogrp**). The need for this last table is predicated on the requirement that clinical findings and other results be linked to a specific subject group receiving a specific exposure regimen. It is possible that one subject group would receive more than one exposure regimen and demonstrate different findings as a result. The contents of these three tables are described in Appendix E. All three tables are linked to a specific study design by means of the citation and design numbers assigned when the design is entered into the database.

III.E.2.b.i. Subject Group Data

When the user chooses **Subject** from the menu, s/he is prompted to enter the study design number which includes the subject data to be entered. After entering this number, s/he then sees the screen depicted in Figure 10 below. The program looks up the number of subject groups in the design and displays this number in the third field of the screen. As subject groups are entered, the program automatically changes the Group value in an incremental fashion using letters **A** to **Z**. In the figure below, the user sees **Group A of 3 of Design 1**, meaning this is the first subject group of a total of three subject groups in study design number **1**. The user enters the subject data in the appropriate fields. The last field is for the user to indicate the total number of exposure regimens received by this particular subject group in the course of the study.

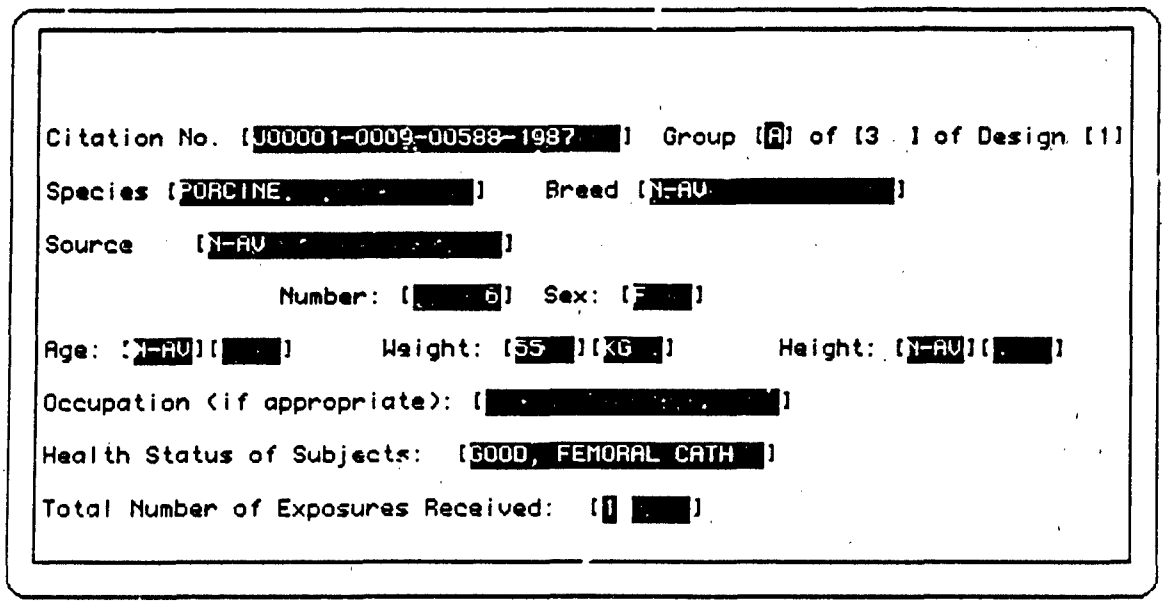


Figure 10. Subject Group Data Entry Screen

III.E.2.b.ii. Exposure Regimens

The second choice on the "Subject - Regimen - Links" menu accesses the Exposure Regimen entry screen. This table holds the data about the agents and regimens the subjects received. The screen shown below in Figure 11 uses some of the same user-oriented enhancements mentioned in regards to the controlling technique entry screen.

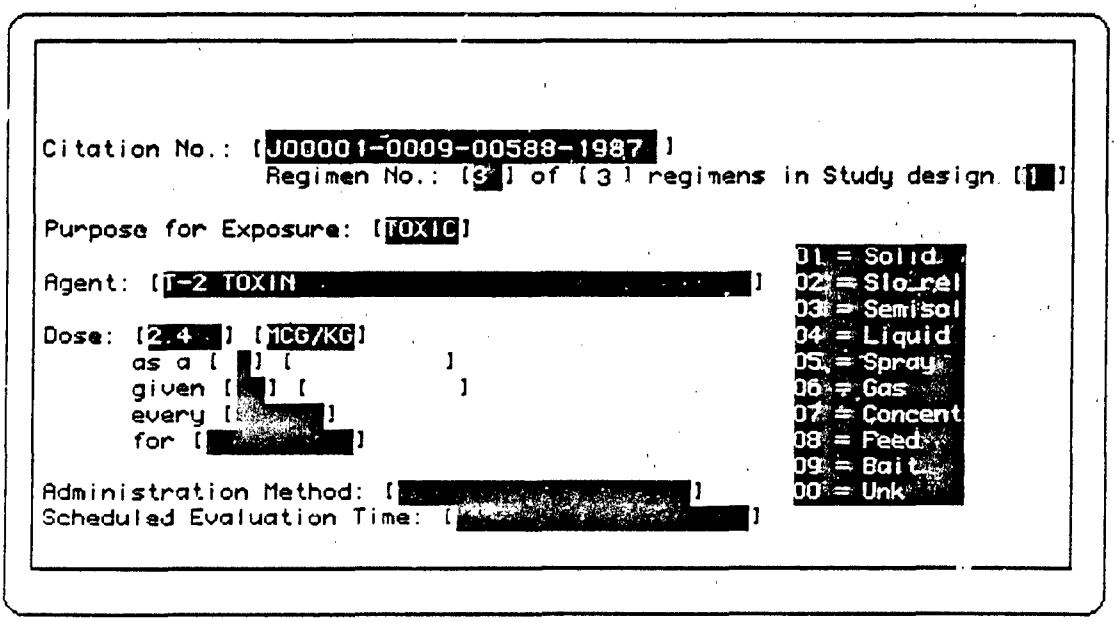


Figure 11. Exposure Regimen Data Entry Screen

In a manner similar to the subject group data, the program looks up the total number of exposure regimens in this design from the **stdydsgn** table and displays this value on the screen. As each individual exposure regimen is entered, the program assigns an exposure number in increments of **01** to **99**. After the linking and sequencing information is attended to, the user indicates the purpose for the exposure. In the paper shown, this particular regimen was given to produce toxic effects. Treatment regimens are also entered this way. After indicating the purpose, the user indicates the specific agent used in the exposure. Eventually, this will be linked to a generic agent controlled vocabulary to assure correctness and consistency in entering this data.

The user next enters the amount of the agent the animal received and the units for measuring the amount. We plan to use a conversion process in the future to convert all entered doses into milligram/kilogram units for consistency. As the user continues, the available codes appear in the right side of the screen. This allows the use of codes for compactness and ease of sorting, and yet the user can see possible choices on the screen. As the codes are entered, the meaning of the code appears in the field next to it. The user enters the dose, dosage units, the formulation, the route of administration, the interval between doses, and the duration or number of doses.

III.E.2.b.iii. Exposure Group Data

After the subject group data and exposure regimen data are entered, the user then needs to create the links between these two information sets. By selecting the "Links" menu option, the user will see the screen represented in Figure 12 below.

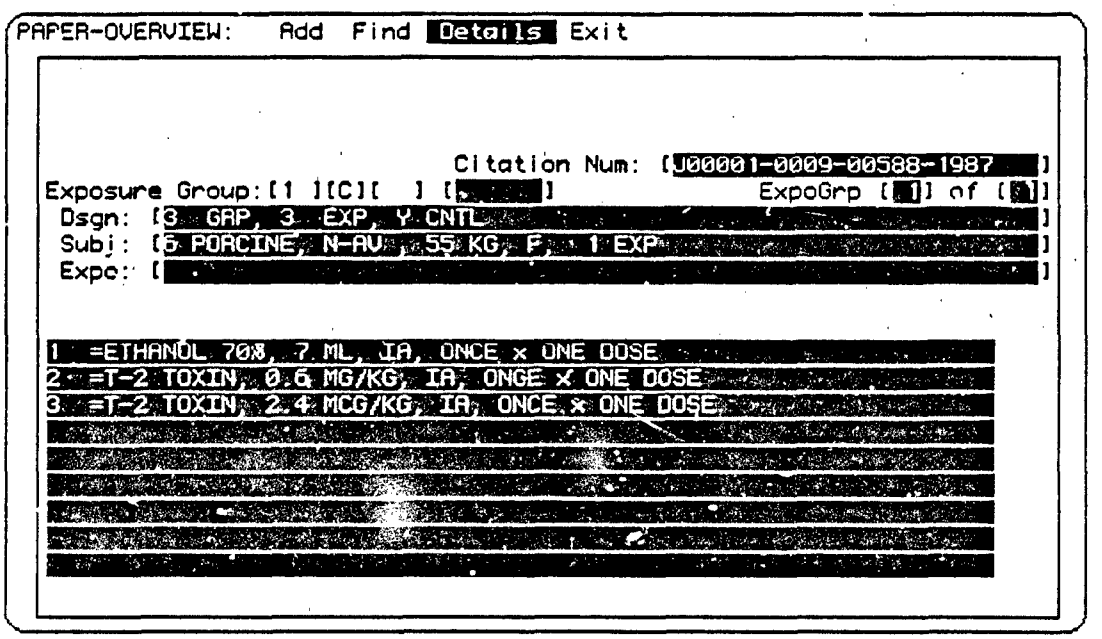


Figure 12. Exposure-Group Link Creation Screen

The purpose of this database table (which is described in Appendix E) is to hold the design-subject group-exposure regimen link which will be used to connect this

information to the results data. In general, the contents consist of the link code and brief descriptions of the study design, the subject group, and the exposure regimen. The descriptions will serve at least two purposes: to give on-screen verification of this information when the results data are being entered, and to provide this information as needed in the structured monographs.

The program presents the user with options based on previously entered data and on his/her current choices. The user first confirms the citation number and is then presented with descriptions of all the study designs entered for this paper. The program creates the descriptions by extracting data from the **stdydsgn** table. After the user selects the desired study design number, the program puts the selected study design description in the appropriate field and then presents computer-prepared descriptions of the subjects involved in the selected design. Choosing the subject group results in the selected subject description being displayed, followed by presentation of similar descriptions for all exposure regimens in this design. In Figure 12, the user has selected design number 1, which is a controlled study involving 3 subject groups and 3 exposure regimens. Subject group C was selected. The corresponding description, which indicates that the group consisted of 6 female pigs with an average weight of 55 kg and were exposed once, was put into the subject group description field. The age of these pigs was not presented in the paper, thus the **N-AV** in the description. The user must now choose between the three exposure regimen options displayed on-screen. From the paper, the user knows that group **C** received exposure regimen **3** and would enter this number. After the exposure option is chosen, the exposure-group link will be created, which in this case would be **1.C03**, meaning design **1**, group **C**, and exposure **03**. This number would be difficult to use without the descriptions stored with the link.

III.E.3. Results Section

The results section was not fully developed in the first year. The primary obstacle has been the selection and/or development of a controlled vocabulary for clinical findings. Our difficulties in deciding on such a vocabulary are discussed in detail below. The importance of this vocabulary cannot be overstated. If clinical findings from a wide variety of papers are to be compiled, the terms that have been entered must adhere to certain rules. We plan to arrange the clinical findings in the structured monograph by the body organ system, followed by the organ, and then the specific clinical finding. To enter this detailed data for each clinical finding in the most efficient fashion, we will use a sign code which will look up the detailed information from the clinical findings controlled vocabulary and insert the needed information in the clinical findings entry screen. This will require that the clinical findings vocabulary be in place when the clinical findings entry program is being tested.

The development of the exposure-group link was necessitated by the clinical findings entry process. By entering the exposure-group link along with a clinical finding, the association between a group's exposure to an agent and the resulting effects is established. To enter the clinical finding data, at least two different data entry mechanisms will be used. One mechanism will focus on one clinical finding and the many associated exposure groups. This will be especially useful for entering tabular data. The other mechanism will focus on one exposure group

and the many associated clinical findings. This is more likely to be used for textual data. These two means of clinical finding entry are shown in Appendix G.

III.E.4. Discussion and Comments Section

The most difficult problem we face in designing the Toxin Knowledge System is how to collect and represent the authors' discussion and others' comments on the authors' work in a compilable form. As authors report their work, they discuss the impact their work has on the understanding of the problem being studied. They frequently review and critique previous research and comment on how their work compared to the previous work. This section of a scientific paper is especially critical with metabolism and mechanism of action studies.

We plan to use separate tables to hold metabolism and mechanism data. The use of the tables in the abstracting process will be determined by the study design type. The authors' observations about each of these areas will be collected and available for compilation. We expect to define general terms for sorting metabolism and mechanism data, thus permitting the compilation of this information in the structured monographs.

The authors' comments about another paper will be recorded in a comments table. This table will include the Toxin Knowledge System abstractor comments, authors' comments about other papers, and commentary from editorials and letters. The use of these comments in the structured monograph has not been defined at this time. One possible use is as annotations in the bibliography.

III.F. Controlled Vocabulary Difficulties

III.F.1. Clinical Finding Controlled Vocabulary

We had hoped that we might be able to directly use an existing vocabulary for clinical findings. We gave consideration to three such vocabularies: the National Library of Medicine *Medical Subject Headings*, the World Health Organization *International Classification of Diseases*, and the American College of Pathology *SNOMED* and associated American Veterinary Medical Association *SNOVET*.

The *Medical Subject Headings* and the *International Classification of Diseases* were considered to have strength in disease terminology but did not have the specific pathology information we believed necessary for describing clinical findings in published papers. Both of these systems were easily understood and could be used with limited modification Neither system was considered adequate for describing clinical findings in animals, a necessity when describing the results of animal studies. The *International Classification of Diseases* had no specific veterinary terms and the *Medical Subject Headings* only had a limited set of such terms.

The *SNOMED/SNOVET* combination was determined to be the most appropriate foundation for our building a clinical finding vocabulary. This combination has a broad set of specific pathologic terms applicable to both human and animal settings. Unfortunately, the *SNOMED/SNOVET* coding system is based on a multiple axis arrangement which can result in a user having to enter from 3 to 6 code numbers to characterize one clinical finding. We considered this to be inappropriate for our intended use. Instead, we decided to use the *SNOMED/SNOVET* system as a foundation for building a clinical finding vocabulary that will use the single-axis terms where possible and integrate the multi-axis terms into a single term. The *SNOMED/SNOVET* computer tapes have

been obtained and are on-line at this time. The process of building the Toxin Knowledge System clinical finding vocabulary is continuing.

A beneficial off-shoot of having the *SNOMED/SNOVET* terms available is the possible use of the Topography terms for sites of administration in the exposure regimen data. Frequently, it is important to know precisely where in the body the researchers administered the toxic or therapeutic agent. The use of a specific topography terminology appears to be a possible solution.

III.F.2. Chemical Name Controlled Vocabulary

The controlled vocabulary for chemical names was expected to be the most straightforward of the controlled vocabularies, as we intended to use the *Registry of Toxic Effects of Chemical Substances* (RTECS) prepared by NIOSH. This registry contains the type of information we believed useful to the Toxin Knowledge System. Our goal was to have the RTECS data on-line as a look-up system for the toxic and therapeutic agents entered in the exposure regimen. When we received the current RTECS computer tape, we found that it contained over 150 megabytes of data. Bringing this up as a relational database system with appropriate indexes for good search performance was estimated to require over 300 megabytes of storage space. The current disk drive configuration of the Sequent™ minicomputer does not have this much contiguous space available. For the RTECS to be used as we originally conceived, it must be a part of the Toxin Knowledge System database files. This would require more contiguous disk space than we have available in any configuration. Our current plans involve extracting and placing data regarding low molecular weight toxins and selected pharmaceutical agents from the RTECS tapes into a file within the Toxin Knowledge System database. We would use this file as the vocabulary for the exposure regimen table and thus provide on-line queries and spelling checks.

III.G. Treatment Database Begun

We anticipated that the treatment monographs would need to be prepared differently. Early in the process, we decided to design a treatment database to provide a clear picture of the information we would need to have as output from the other sections. While we were learning Informix-4GL™ programming techniques, such a database was created. This database design has influenced some aspects of the overall design, especially in how treatment regimens will be handled in the abstracting process. The complexity of the inter-table relationships prevented our using Informix-SQL™ data entry techniques. Informix-4GL™ programs for this have not been written, as we do not intend to keep the treatment information as a separate database. Instead, this information will be incorporated within the full Toxin Knowledge System.

V. CONCLUSIONS

Toxin Knowledge System development is well underway with major database tables in place and the data manipulation programs operational. The use of Informix-4GL™ has permitted the creation of a sophisticated user interface which permits smooth and consistent data entry. This programming language is based on a high performance relational database management system, Informix-SQL™.

The foundation of any literature-based system is the citation information. Citation management functions for the Toxin Knowledge System are fully

operational with well over 1600 toxin and toxin research-related citations entered. The database tables and the corresponding program modules for this section function well. Journal titles and abbreviations are controlled by means of a separate database table containing entries for over 6000 journals.

The tables involving the content of the paper itself are partially complete. Tables and program modules exist for paper overview, study design, subject groups, and exposure regimens. A table with data derived from the study design, subject groups, and exposure regimens has also been developed. These require further testing with entry of more papers to identify areas needing refinement. The user-interface for the paper content sections has not been extensively tested, and we anticipate several modifications will be required to achieve the needed user-program interactions.

Clinical finding data is currently being analyzed for inclusion in the system. This area is pivotal in the success of this system. Part of the difficulty in this area is the establishment of a controlled vocabulary for clinical findings. Preliminary work on using *SNOMED/SNOVET* as the foundation of this controlled vocabulary has begun. As soon as the database table design is resolved for both the controlled vocabulary and the reported clinical findings, the programs for data entry and manipulation will be created.

In addition, other database tables need to be designed and the corresponding program modules written. The most important of these are the chemical controlled vocabulary, analytical methods tables, and analytical results table.

Structured monograph generation is a critical element which has been considered but not directly addressed. This has been largely due to the requirement that the database tables be completed first. When the clinical findings tables are in place the work on this element can begin.

Treatment information issues have been studied and a separate treatment database was initially created. This process was informative but this separate database will eliminated eventually. Rather, the integration of treatment information into the overall Toxin Knowledge System is considered to be the best means of making the information available.

VI. RECOMMENDATIONS

The development of the Toxin Knowledge System should be continued, as the benefits of a structured toxin information gathering system are now apparent. We recommend that the development process focus on managing core information. This broadly involves citation data, study design, analytical methods, subject groups, exposure regimens (including treatment modalities), results (including clinical findings and analytical results), mechanisms, pharmacokinetic data, comments, and structured monograph generation. This information management should address database table design, the user-interface for abstracting data, and programs to manipulate the collected data. The current collection of toxin research papers should serve as an initial source of information but mechanisms to add more current papers should be implemented as well. It is further recommended that the following areas be made lower priority: keyword management, study evaluation, statistics methodology, and management of actual graphic data, such as micrographs.

VII. APPENDICES

Appendix A

Content of Citation Section Database Tables

Journlst

Booklst

Citation

Authors

journlst

The purpose of this table is to provide a controlled listing of journals to be used in the citation process. This will permit journals and book citations to be entered in a similar fashion while maintaining the unique aspects of each citation form in their repsective tables.

jaquis - serial

Serially assigned number for each journal in the system

jcode - char(20)

A unique code for each journal in system composed of **J** and the jacquis number. While this number is 20 characters long in the database table, only 6 characters are actually used. The 20 characters are necessary to join the serial table and and the **J** together.

jname - char(120)

The exact name of the journal. Most are taken from the List of Journals Indexed by NLM.

jabrv - char(50)

Journal abbreviation, generally taken from List of Journals Indexed by NLM. These will be used in the reference listings and to display on screen when a journal code is entered in the citation table.

booklst

The purpose of this table is to provide a controlled listing of books to be used in the citation process. This will permit journals and book citations to be entered in a a similar fashion while maintaining the unique aspects of each citation form in their repsective tables.

bacquis - serial
Serially assigned number for each book in the system

bcode - char(20)
A unique code for each book in system composed of **B** and the bacquis number. While this number is 20 characters long in the database table, only 3 characters are actually used. The 20 characters are necessary to join the serial table and and the **B** together.

bname - char(60)
The actual title of the book

bedno - char(2)
The edition number of the book

bvol - char(2)
The volume number of the book

bdate - char(4)
The year of the book's publication: should be edition specific.

bpub - char(20)
Publisher of this edition of the book

bpubplace - char(20)
Place of publication of this edition

beditor - char(50)
The editors of this edition of the book or the author(s) if not an edited work. This is a simple string and is not intended to do any more than complete the citation in a bibliography, etc.

bisbn - char(20)
The is the ISBN number for the book. This may be dropped in the future.

citation

The citation building block of the whole system. Generates a citation number which serves as the primary connector for all other tables. This holds either journal article or book chapter data.

citnumb - char(25)

The relation between all tables. The number will have the following format:

JBJBJB-VVVV-PPPP-YYYY where:

JBJBJB = journal or book code (journlst.jcode or booklst.bcode)

VVVV = journal volume number or book chapter number

PPPPP = first page number

YYYY = year of publication

This number will be generated from data entered in the other columns in the citation table.

citsource - char(20)

Source of the citation; the corresponding journlst.jcode or booklst.bcode

citvol - char(4)

The journal volume number or book chapter number

citpage - char(11)

The inclusive page numbers **PPPPP-PPPPP**

citdate - char(4)

The year of publication

cittitle - char(250)

The actual title of the paper or chapter

citlocate - char(5)

The location of the actual paper in filing systems within the group

citfile - char(12)

A filing system number for the paper. Used to manage actual paper filing process. Currently consists of first 4 characters of first author's last name, the volume number, the first page number, and the last 2 digits of year.

authors

The purpose of this table is to hold author data for each paper or chapter entered into the system.

aucitnumb - char(25)
The link to citation.citnumb

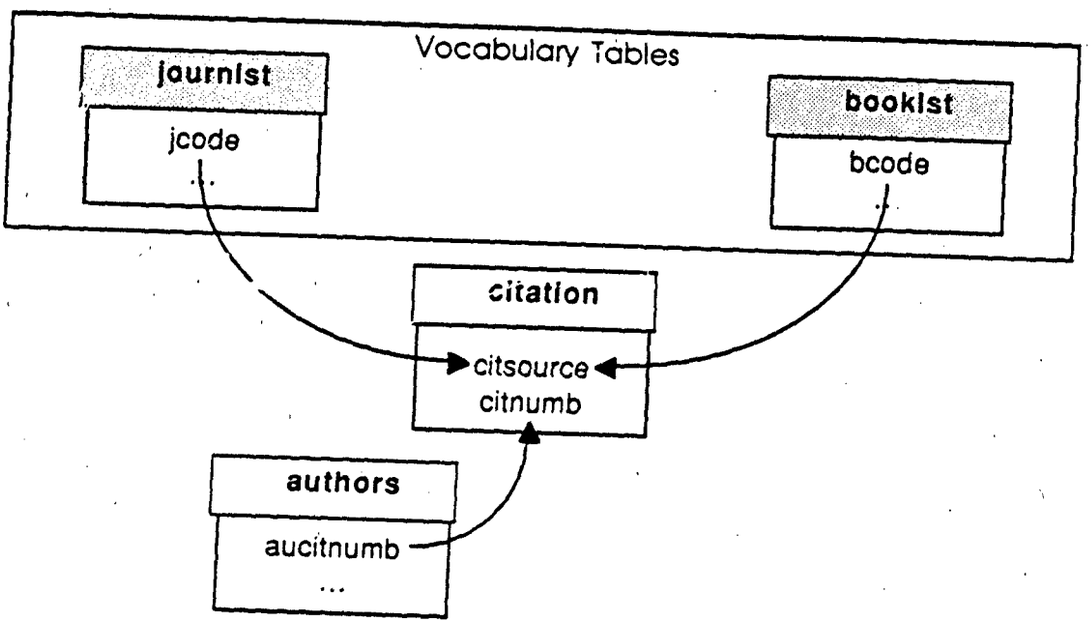
aucitfile - char(12)
Link to citation.citfile

authname - char(50)
The name of the author formated as follows: last name, space, initials. No punctuation is to be used.

authsig - char(2)
Publication order for the authors names

Apoendix B.

Interaction of Citation Section Database Tables


Vocabulary Tables
journlst
jcode
...
booklst
bcode
...
citation
citsource
citnumb
authors
aucitnumb
...

Appendix C.

Contents of Keyword-Related Database Tables

Keywords

Keylist

keywords

The purpose of this table is to permit searching for unabstracted citations entered into the system. These will also be used to select citations for abstracting.

keycitnumb - char(25)
The link to citation.citnumb

keycitfile - char(12)
Link to citation.citfile

keyword - char(20)
The keyword describing some aspect of the paper, matches the keylist.kword

keycode - char(10)
The keycode which can be used for group look-ups and is used to automatically insert the keyword when the code is entered. This is linked to the controlled vocabulary keylist.kcode

keylist

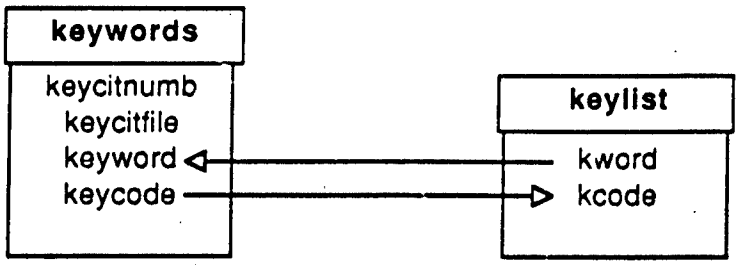
The purpose of this table is to provide a controlled vocabulary for keyword entry into the system

kcode - char(10)
Code number for linking the keyword. User can enter a kcode and the kword will pop up on screen. This can also be used to query for a group of keywords of the same group.

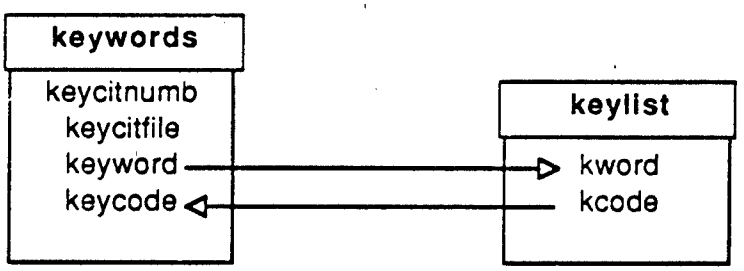
kword - char(20)
The controlled keyword vocabulary. These are arranged by group.

Appendix D.

Interaction of Keyword-Related Database Tables



In this example, the user enters a keycode in the keyword entry screen. The computer program looks up the corresponding keyword in the keylist table and inserts it into the keywords entry screen.



In this example, the user enters a keyword in the keyword entry screen. The computer program looks up the corresponding keycode in the keylist table and inserts it into the keywords entry screen.

Appendix E. Content of Paper-Data Database Tables

Paperover

Stdydsgn

Subjgrp

Exporegm

Expogrp

paperover

The purpose of this table is to hold certain basic information about the paper. It holds the number of study designs within the paper, as well as the purpose for the paper.

papcitnumb - char(25)
The link to citation.citnumb

papcitfile - char(12)
Link to citation.citfile

papstatepur - char(50)
The stated purpose of the paper. This is both an evaluation point and provides information necessary to classify the paper.

papimppur - char(50)
The implied purpose of the paper. This can give insight into biases as well as "the real reason" for the study.

papaim - char(3)
A broad term to describe the aim of the study. Will eventually be used to control abstracting process flow.

papnumdsgn - char(2)
The number of study designs in the paper

stdydsgn

The purpose of this table is to hold certain basic information about the study. It holds confirming information about the number of groups involved, the number of exposures involved, and the presence or absence of controls. Assigned a number from 1 to 99 to identify this study within the paper describing it.

stycitnumb - char(25)
The link to citation.citnumb

stycitfile - char(12)
Link to citation.citfile

stydsgncur - smallint
A number to identify this design from others in the paper. Used to link to subjgrp and exporegm. Also used in creation of expogrp.eglink.

stydsgntot - smallint
The total number of study designs in the paper. Linked to paperover.papnumdsgn.

stytype - char(2)
The broad type of study. This will be used to further control the abstracting process.

styvivvit - char(1)
Indication of whether the paper describes an in vivo or an in vitro experiment.

stynumgrp - char(2)
The number of different subject groups studied

stynumexp - char(2)
The number of different exposure regimens used

stycntl - char(1)
Flag to whether or not controls were used.

stycntlcmp - char(2)
The group comparison information. (Within group, between groups, combination)

stycmpmeth - char(20)
The method for comparing the groups regardless of within or between

stycntlmeth - char(1)
Control methodology base — concurrent vs non-concurrent

stycntltyp - char(20)
The type of control used for the respective method

stycntassgn - char(20)
The method for assigning the subjects to the group

exporegm

This table holds data on the exposure regimens that the subjects will undergo. Each regimen will be given a number between 00 and 99. This number will be used with the subject group number and study design number to form an exposure-group link.

excitnumb - char(25)
The link to citation.citnumb

exdsgnnum – smallint
The link to the identifying number of the study design.

exlink - char(2)
The link to expogrp.eglink, numbers 00 to 99.

expurpose - char(5)
The purpose of the exposure. For example, toxicity, treatment, or control.

exagent - char(40)
Agent in exposure regimen

exdose - char(5)
Dose of agent used (no units)

exdoseunit - char(6)
Units of dose administered

exformul - char(2)
Formulation of the agent used in the regimen. Formulations will be abbreviated and an abbreviation list will be maintained.

exroute - char(2)
Route of administering the agent in question. Routes will be abbreviated and an abbreviation list will be maintained.

exinterval - char(6)
The interval between multiple exposures, e.g. every 4 hours

exduration - char(10)
The duration of exposure to include both duration of contact as well as number of doses received

exadminmeth - char(20)
The method of administering the agent to the subjects. Not to be confused with route. Example: slow IV via pump. IV is the route, "slow, via pump" is the administration method

exevaltime - char(20)

The time for evaluation; can be interval of evaluation if needed. This particular item may be better maintained in another table, such as study design.

subjgrp

This table holds data about each group of subjects in the study. Each group is assigned a letter from **A** to **Z** sequentially. This letter will be joined with the exposure regimen and study design numbers to create an exposure-group link.

sgcitnumb - char(25)
The link to citation.citnumb

sgdsgnnum – smallint
The link to the identifying number of the study design

sglink - char(1)
Character from **A** to **Z**; used in association with the exporegm.exlink and stdydsgn.stydsgncur to form expogrp.eglink

sgspecies - char(20)
The species of the subjects used; not necessarily the Latin name

sgbreed - char(20)
The breed, race, ethnic, or other genetic variation

sgsource - char(20)
Source of subjects used in study

sgnumb - smallint
Number of subjects in group

sgage - char(4)
The age of the subjects

sgageunit - char(4)
The units for the age of the subjects

sgwt - char(4)
The weight of the subjects

sgwtunit - char(4)
The units for the weight of the subjects

sqht - char(4)
The height of the subjects. This likely to be useful only in human studies for the determination of surface area.

sghtunit - char(4)
The units for the height of the subjects

sgsex - char(4)
The sex of the subject. Use abbreviations

sgoccup - char(20)
The occupation of the subjects; aimed at human subjects

sghlthstat - char(20)
The health status of the subjects. Can include vaccinations, preexisting illnesses, etc

sgtotexpo - smallint
The number of exposures this group received during the study

expogrp

This table holds the links and brief description of the group and exposure regimen. This will be used to link results to the subjects and regimens.

egcitnumb – char(25)
The link to citation.citnumb

egtotnum – smallint
The total number of exposure group links that have been made for this design.

eglink – char(6)
This is comprised of the stdydsgn.stydsgncur (**1** to **99**), a ".", the subjgrp.sglink (**A** to **Z**), and the exporegm.exlink (**00** to **99**). The result would look like **1.A01**. This will be created during the data entry and selection process, and will be used to link the subject group and exposure regimen to a given result.

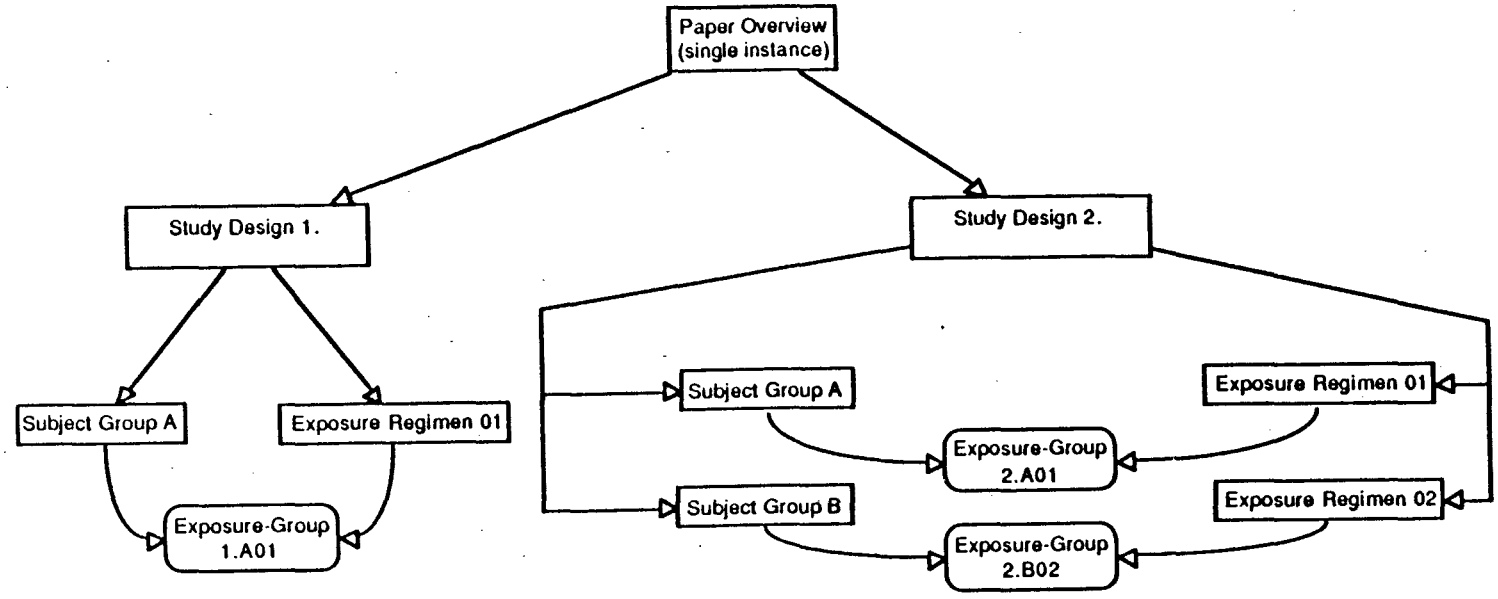
egdsgndsc – char(60)
Brief description of the study design. This is used for on-screen confirmation of the design data when associated result data are entered. This should be generated by the computer and inserted when the user selects the study design data.

egsubgdsc – char(60)
Brief description of the subject group. This is used for on-screen confirmation of the group data when associated result data are entered. This should be generated by the computer and inserted when the user selects the exposure regimen data.

egexpodsc – char(60)
Brief description of the exposure regimen. This is used for on-screen confirmation of the exposure data when associated result data are entered. This should be generated by the computer and inserted when the user selects the exposure regimen data.

Appendix F.

Interactions of Paper Overview Table with the Materials and Methods Tables



Each paper can have only one paper overview entry. The paper can have any number of study designs. These are assigned number codes starting with "1". The code is simply an identifying number. Each study design would be expected to consist of one or more subject groups and one or more exposure regimens. The subject groups are assigned alphabetical codes from "A" to "Z". The exposure regimens are assigned numerical codes from "00" to "01". The user creates the exposure group links by selecting the subject groups and the associated exposure regimens. The linkage created would consist of the study design code, a period, the subject group code, and the exposure regimen code.

Appendix G

Two Means for
Clinical Finding Data Entry

clinical finding XYZ				
expogrp 1.A01	Severity = 100 iu	Frequency 1 / 1	Onset 1 hour	Duration 8 hours
expogrp 2.A01	Severity = 200 iu	Frequency 9 / 10	Onset .25 hour	Duration 8 hours
expogrp	Severity = 1 iu	Frequency 1 / 10	Onset 4 hours	Duration .25 hour

This shows how the exposure group links are used to indicate which subjects receiving which exposure regimen showed the specified clinical finding. This method of entry is especially useful for tabular data.

expogrp 2.A01				
clinical finding ABC	Severity = 3+	Frequency 8 /10	Onset 1 hour	Duration 8 hours
clinical finding XYZ	Severity = 200 iu	Frequency 9 / 10	Onset .25 hour	Duration 8 hours
clinical finding MNO	Severity = 1	Frequency 6 / 10	Onset 1.5 hours	Duration 12 hour

This shows how the exposure group links are used to indicate the clinical findings seen with a specified group of subjects receiving a specified exposure regimen. This method of entry is especially useful for textual data in which the author indicates that "the group receiving 'such and such' showed 'sign A', 'sign B,...'".

VIII. DISTRIBUTION LIST

5 copies	Commander US Army Medical Research Institute of Infectious Diseases ATTN: SGRD-UIZ-M Fort Detrick, Frederick, MD 21701-5012
1 copy	Commander US Army Medical Research and Development Command ATTN: SGRD-RMI-S Fort Detrick, Frederick, MD 21701-5012
2 copies	Defense Technical Information Center (DTIC) ATTN: DTIC-DDAC Cameron Station Alexandria, VA 22304-6145
1 copy	Dean School of Medicine Uniformed Services University of the Health Sciences 4301 Jones Bridge Road Bethesda, MD 20814-4799
1 copy	Commandant Academy of Health Sciences, US Army ATTN: AHS-CDM For Sam Houston, TX 78234-6100